Prominent Users Detection during Specific Events by Learning On- and Off-topic Features of User Activities

Imen Bizid*†, Nibal Nayef*, Patrice Boursier*‡‡, Sami Faiz†, and Jacques Morcos*
*L3i, University of La Rochelle, Avenue Michel Crépeau, 17042, La Rochelle, France
Emails: {imen.bizid, nibal.nayef, patrice.boursier, jacques.morcos}@univ-lr.fr
†LTSIRS, University of Tunis, 1005, Tunis, Tunisie
Emails: {sami.faiz}@insat.rnu.tn
‡IUMW, City Campus, Jalan Tun Ismail, 50480 Kuala Lumpur, Malaysia
Emails: {patrice}@iumw.edu.my

***Abstract*—Microblogs such as Twitter are characterized by the richness and recency of information shared by their users during major events. However, it is very challenging to automatically mine for information or for users sharing certain information due to the huge variety of unstructured stream of data shared in such microblogs. This work proposes a ranking and classification model for identifying users sharing useful information during a specified event. The model is based on a novel set of features that can be computed in real time. These features are designed such that they take into account both the on and off-topic activities of a user. Once users are characterized by a feature vector, supervised machine learning tool is trained to classify users as either prominent or not. Our model has been tested on data shared during a flooding disaster event and performed very well. The achieved results show the effectiveness of the proposed model for both the classification and ranking of prominent users in such events, and also the importance of the adjustment of the on-topic features by the off-topic ones when describing users' activities.**

Keywords*—Key user identification, On- and off-topic features, Events in Twitter, SVM Classification and Ranking, Real-time feature computation.

I. Introduction

Microblogging platforms offer the services of convenient access to and sharing of fresh data on any topic. Shared data is usually limited to a specified number of characters, for example, tweets are limited to 140 characters. This results in a huge stream of unstructured data, which makes information retrieval within such data very challenging, specially when having to perform such tasks in real time.

Having the aforementioned particularities of microblogs in mind, many research works have associated the relevance and the quality of the shared messages in microblogs with the user's prominence in terms of the network and the targeted topic but not in terms of messages' content [1], [2].

Those work have focused on the identification of influencers [3], [4] and domain experts [5], [6] in microblogs. The identification of the latter type of users is usually based either on their centrality and popularity in the network or on their frequent networking activities domains. From another aspect, domain experts are generally identified by analyzing their historic information regarding a topic of interest. However, the prominence of microblogs users during a *specific event* cannot be evaluated according to the user's centrality or prior activities in the network.

The features characterizing prominent users should reflect the nature of a user's behavior during such events. Consider the case of a disaster where alerts and emergency states are being rapidly updated over time. The prominent users – whom we are interested in – would focus their attention and communication mostly on the topical information related to the disaster. These users are not necessarily experts in disasters, they may be ordinary microblogs users geolocated in the disaster area and who are sharing what they are seeing and experiencing. Hence, they would share a lot of exclusive information. While users such as media channels toggling between several topics would share many event-related information which are already diffused in the network. Thus, features based on traditional metrics such as the number of shared on-topic tweets cannot be directly used to identify user's prominence. Therefore, identification models which are based on traditional metrics would be sensitive to users interested in several topics and sharing outdated information.

In this paper, we propose a model that alleviates these shortcomings first by designing a set of metrics which represent Twitter users interested in a specific event. These metrics characterize each active user by considering both the on- and off-topic activities of the user during the event. Based on those metrics, our list of features promote users who focus only on the event under consideration, and penalize those who are toggling among several topics. Using these representative features, we use supervised learning to train an SVM model to identify the most prominent users in real time during an event.

The rest of this paper is organized as follows: Section II reviews related work. Section III presents the set of our proposed features used to model microblogs' users. Section IV summarizes the classification and ranking approach employed to identify and detect the best prominent users. Section V presents the experiments and results obtained by our model. Section VI concludes the paper and discusses future work.

II. Related Work

To the best of our knowledge, the issue of prominent users' identification has never been explored in the context of specific events. However, there have been several attempts proposing